Predicción de Series Temporales: Red de Transporte Público de Madrid

Ramón Morillo Barrera

Análisis y Predicción de la Demanda de Transporte Público en la Comunidad de Madrid

Propósito del Informe

Este informe aborda la necesidad de analizar y predecir la demanda de transporte público en la Comunidad de Madrid, una región clave con una alta densidad poblacional y un sistema de transporte esencial para la movilidad diaria de millones de personas. A partir de datos históricos de viajeros entre enero de 2012 y septiembre de 2024, se busca construir un modelo robusto que permita tomar decisiones informadas para la planificación, mejora y sostenibilidad del sistema.

Objetivos

1. **Análisis Histórico**: Examinar la evolución del número de viajeros para identificar patrones como estacionalidad, tendencias y ciclos.
2. **Comparación de Modelos**: Evaluar distintos enfoques predictivos, desde modelos estadísticos clásicos hasta técnicas avanzadas de machine learning.
3. **Predicciones Confiables**: Estimar la demanda de pasajeros para el resto de 2024 y todo 2025, tanto a nivel mensual como trimestral.
4. **Apoyo a la Toma de Decisiones**: Proveer información útil para la planificación de políticas públicas y asignación de recursos.

Metodología

Se aplicaron múltiples enfoques predictivos, incluyendo:

- **Modelos estadísticos**: ARIMA, ETS, BATS, TBATS, CES, Prophet entre otros.
- **Técnicas de Machine Learning**: Decision Tree, Random Forest, XGBoost, LightGBM entre otros.

Cada modelo fue ajustado y evaluado mediante métricas clave como MSE, RMSE y MAPE, utilizando técnicas de validación cruzada y ajuste de hiperparámetros.

Hallazgos Clave

- **Modelos Destacados**: Los modelos ETS y Theta presentaron el mejor desempeño en términos de precisión y ajuste a los datos históricos.
- **Recomendaciones**: Basar las predicciones futuras en el modelo ETS, que demostró mayor capacidad interpretativa y robustez.

Impacto Esperado

Las predicciones generadas servirán como insumo para diseñar estrategias de transporte más eficientes, adaptadas a la evolución de la demanda y alineadas con los objetivos de sostenibilidad y mejora de la experiencia del usuario.

Resumen Ejecutivo

La Comunidad de Madrid, con su vibrante dinamismo y alta densidad poblacional, depende en gran medida de su infraestructura de transporte público para garantizar la movilidad diaria de millones de ciudadanos. En un contexto marcado por el crecimiento poblacional y la recuperación tras la pandemia de COVID-19, predecir la demanda de viajeros en el sistema de metro y autobuses se ha convertido en una prioridad estratégica. Este análisis no solo busca optimizar recursos y mejorar la calidad del servicio, sino también asegurar la sostenibilidad del sistema a largo plazo.

Para abordar esta necesidad, se ha analizado el archivo **Madrid.csv**, que recoge datos mensuales del número de viajeros transportados en la región desde enero de 2012 hasta septiembre de 2024. A partir de esta información, se han planteado dos objetivos principales: crear series temporales que reflejan la evolución histórica de los viajeros y generar predicciones confiables para el resto de 2024 y 2025. Estas predicciones no solo informarán decisiones de planificación, sino que también ayudarán a detectar patrones clave, como estacionalidad y tendencias, que afectan directamente la operativa del sistema.

El enfoque metodológico ha sido integral y riguroso, integrando tanto modelos estadísticos básicos con el objetivo de proporcionar al lector una introducción de la metodología utilizada y el funcionamiento de los modelos estadísticos de predicción como estadísticos avanzados como ARIMA, ETS, BATS, TBATS, CES, BSTS, Theta, y Prophet, así como técnicas de machine learning como Decision Tree, K-Nearest Neighbors, Random Forest, XGBoost y LightGBM. Cada modelo ha sido evaluado en función de métricas de error y capacidad interpretativa, con el objetivo de seleccionar aquel que ofrezca las predicciones más precisas y útiles desde una perspectiva práctica.

Los resultados de este análisis proporcionan información clave para la planificación de políticas públicas y la asignación de recursos, contribuyendo al fortalecimiento de un sistema de transporte público más eficiente, sostenible y adaptado a las necesidades de los ciudadanos. Este informe no solo expone los hallazgos más relevantes, sino que también detalla la lógica detrás de cada decisión técnica y su impacto en los objetivos estratégicos de la Comunidad de Madrid.

Este informe se estructura en varias secciones clave para facilitar la comprensión del análisis. Primero, se presentan los aspectos principales del análisis exploratorio de datos. Luego, se incluye una breve descripción de cada modelo predictivo empleado, junto con sus gráficas de predicción y las métricas de evaluación. Finalmente, se concluye con un análisis que destaca los hallazgos y recomendaciones más relevantes.

Caso de negocio

La Comunidad de Madrid, una de las regiones más densamente pobladas y dinámicas de España, depende de su sistema de transporte público para facilitar la movilidad de millones de personas diariamente. Con el crecimiento constante de la población y el incremento en la actividad tras eventos disruptivos como la pandemia de COVID-19, surge la necesidad de optimizar la planificación y operación de la red de transporte público. Para ello, la predicción precisa de la demanda futura es crítica para garantizar un servicio eficiente, reducir costes operativos y mantener la sostenibilidad del sistema.

El transporte público enfrenta retos significativos debido a fluctuaciones en la demanda y patrones estacionales complejos. Una planificación ineficaz puede derivar en un uso ineficiente de recursos, sobrecostes y deterioro del servicio, lo que impacta tanto a los operadores como a los usuarios. Además, es crucial identificar y contrastar patrones de tendencia y estacionalidad para tomar decisiones informadas.

Objetivos

1. **Predicción de la Demanda**: Utilizar modelos avanzados como ARIMA, ETS, THETA, BATS, entre otros, para predecir la demanda futura de transporte público, basada en datos históricos de viajeros.
2. **Análisis de Patrones**: Contrastar hipótesis relacionadas con la estacionalidad y las tendencias observadas en los datos para comprender mejor los factores que afectan la demanda.
3. **Optimización de Recursos**: Minimizar costes operativos y mejorar la eficiencia de la red de transporte público mediante una planificación basada en datos.

Alcance del Proyecto

Se analizarán los datos mensuales y trimestrales del archivo Madrid.csv, que contienen el número de viajeros transportados en metro y autobuses desde enero de 2012 hasta septiembre de 2024. Los modelos seleccionados se evaluarán en términos de precisión y capacidad predictiva, y se contrastarán los patrones de comportamiento estacional y tendencia detectados.

Beneficios Esperados

- **Eficiencia Operativa**: Reducción de costes a través de la optimización del despliegue de recursos en función de la demanda esperada.
- **Satisfacción del Usuario**: Mejora en la calidad del servicio al prever y abordar adecuadamente picos de demanda.
- **Sostenibilidad**: Implementación de estrategias más sostenibles al alinear recursos con las necesidades reales.

Plan de Acción

1. **Preparación de los Datos**: Limpieza y análisis exploratorio del conjunto de datos Madrid.csv.
2. **Implementación de Modelos**: Entrenamiento de los modelos ARIMA, ETS, THETA, BATS, entre otros, con los datos históricos.
3. **Validación**: Evaluación del desempeño de los modelos mediante métricas de error predictivo (RMSE, MSE y MAPE).
4. **Análisis de Hipótesis**: Identificación y contrastación de patrones estacionales y de tendencia.

Índice

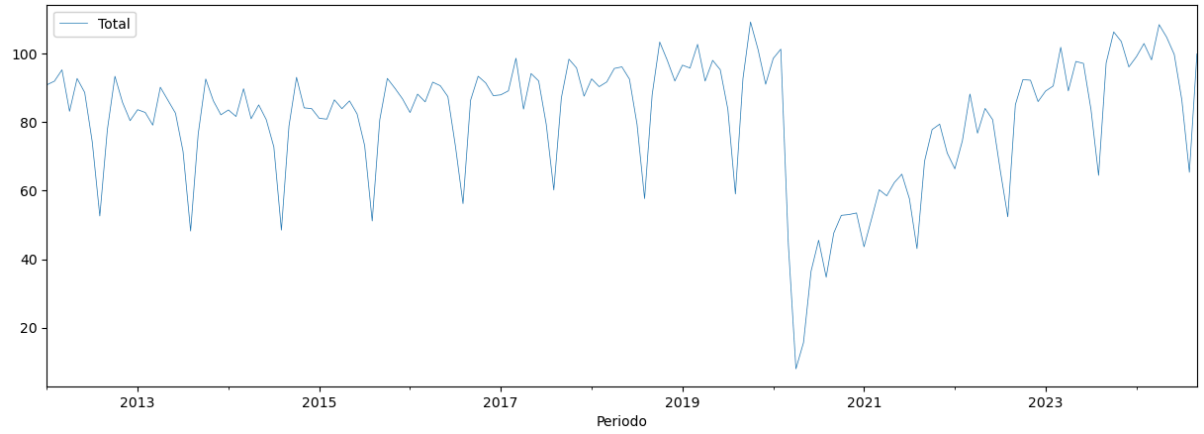
1. Análisis exploratorio de los datos

Antes de comenzar a redactar el procedimiento y las conclusiones extraídas del análisis exploratorio de los datos, me gustaría comentar que he elaborado las predicciones y el graficado de las mismas con datos referentes a miles de pasajeros. Es decir, en el eje Y de las series temporales graficadas se podrá observar como los valores avanzan de 20 en 20 en el caso de las mensuales y de 50 en 50 en el caso de las trimestrales. Esto quiere decir que 100 hace alusión a 100,000 (cien mil) pasajeros.

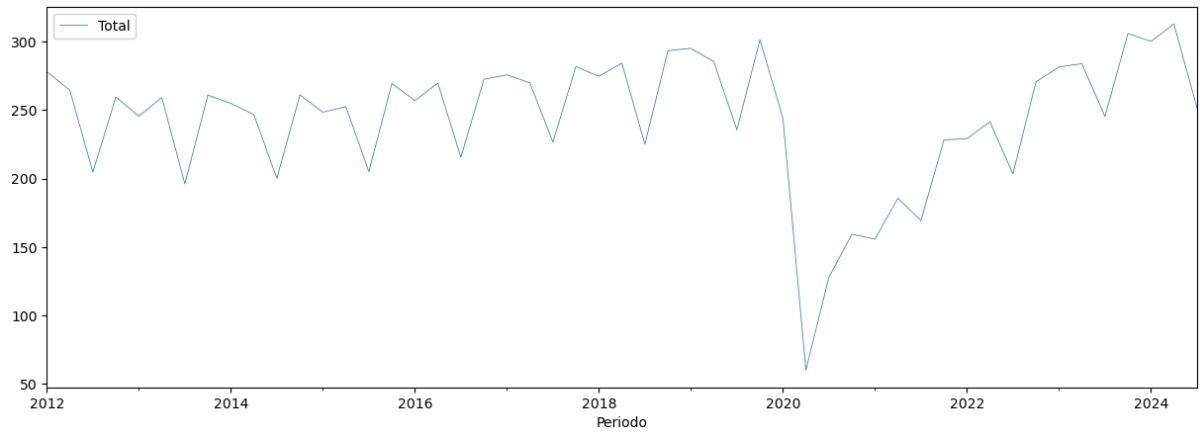
Debido al parecido de las tres series temporales, para no alargar y repetir el contenido del informe, únicamente me centraré en mostrar la serie combinada de pasajeros de autobús y metro de Madrid. He de enfatizar que en los notebooks se han analizado y generado predicciones de las tres series temporales por separado. He tomado esta decisión ya que no he llegado a encontrar motivos de relevancia para analizarlas en el informe por separado debido a su gran parecido en tendencia y estacionalidad.

Una vez aclarado estos detalles, para que no haya confusión alguna, procederé a redactar las características de la serie temporal y de los hallazgos encontrados en el análisis exploratorio de los datos.

1.1 Series temporales originales



La serie temporal mensual refleja una **tendencia creciente** en el número de pasajeros hasta 2019, con un patrón **estacional claro** de picos en primavera y otoño, y caídas en verano y periodos vacacionales. Destaca un **valor atípico en marzo de 2020**, donde se registra una caída abrupta debido a la pandemia de COVID-19 y las restricciones de movilidad. Tras este evento, se observa una recuperación gradual que retoma los patrones previos de tendencia y estacionalidad.

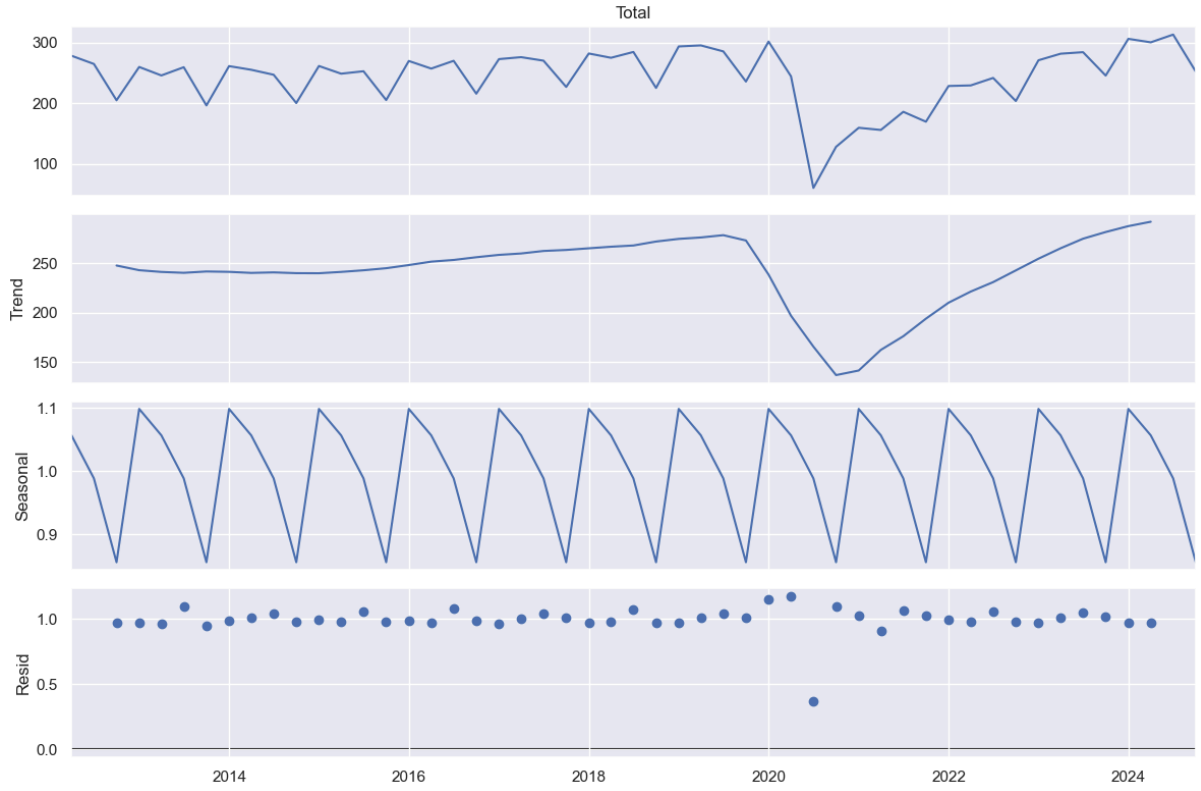


La serie temporal trimestral muestra una **tendencia creciente** en el número de pasajeros hasta 2019, con fluctuaciones estacionales más suaves que reflejan el comportamiento típico del año, destacando trimestres más altos en primavera y otoño. En marzo de 2020 se observa un **valor atípico significativo** con una caída abrupta debido a la pandemia de COVID-19 y las restricciones de movilidad. Posteriormente, se evidencia una recuperación progresiva, retomando los niveles de actividad cercanos a los previos a la pandemia hacia 2023.

1.2 Estacionalidad

Como hemos comentado, se observa una clara estacionalidad, con una tendencia a decrecer en el tercer trimestre del año en la que la actividad económica es menor y los centros educativos cierran debido a la finalización del periodo escolar.

1.3 Descomposición Aditiva



Al realizar la descomposición aditiva de la serie temporal, se observa un ligero crecimiento de la tendencia, que se ve afectado por la pandemia de COVID-19, la cuál se va recuperando hasta alcanzar los valores anteriores a la pandemia a finales del año 2023.

En cuanto al componente estacional observamos lo anteriormente comentado. Existe una clara disminución del uso del transporte público de Madrid en los meses de verano, temporada vacacional.

En cuanto al residuo, los puntos parecen distribuirse de manera consistente en torno a un valor cercano a 1, aunque hay ciertas desviaciones notables (como en 2020, con un residuo significativamente bajo). La pandemia de COVID-19, afectó la demanda de transporte público. Este tipo de residuo fuera de lo normal sugiere que el modelo no pudo prever adecuadamente eventos excepcionales.

1.4 Conclusiones EDA

1. Tanto la serie de autobús como la de metro son muy parecidas aunque el metro cuenta con una mayor cantidad de pasajeros, según mi hipótesis, se debe a la buena conectividad que tiene la red de transporte público del Consorcio Regional de Transportes de Madrid, la cuál proporciona diferentes rutas combinables gracias a los intercambiadores situados en diferentes puntos de la ciudad los cuales permiten que los pasajeros tomen la ruta más corta

combinando ambos medios de transporte. Esto proporciona una mayor afluencia de pasajeros en el tiempo, lo que se traduce en un aumento de los ingresos.

2. Se nota una caída en la demanda de transporte público en el mes de marzo del año 2020 debido al decreto del Estado de alarma causado por la pandemia del virus COVID-19, situación a tener en cuenta a la hora de tomar medidas económicas que gestionen la rápida caída de la demanda con el objetivo de optimizar la eficiencia de la red de transporte público de Madrid, así como los costes ocasionados y los ingresos recibidos.

3. Observamos un claro componente estacional en ambas series en el tercer trimestre del año, según mi hipótesis esta estacionalidad que causa una disminución de la demanda de transporte público se debe al cierre de instituciones educativas con motivo de finalización del curso escolar, baja actividad empresarial y disminución de la población en ese periodo fruto del periodo vacacional el cuál gozan los trabajadores.

Después de esta introducción donde se han generado varias hipótesis y el lector habrá podido comprobar de manera visual la evolución de la serie temporal, la tendencia y el componente estacional, procederé a mostrar algunos de los modelos utilizados para generar predicciones además de las métricas de error establecidas para su evaluación.

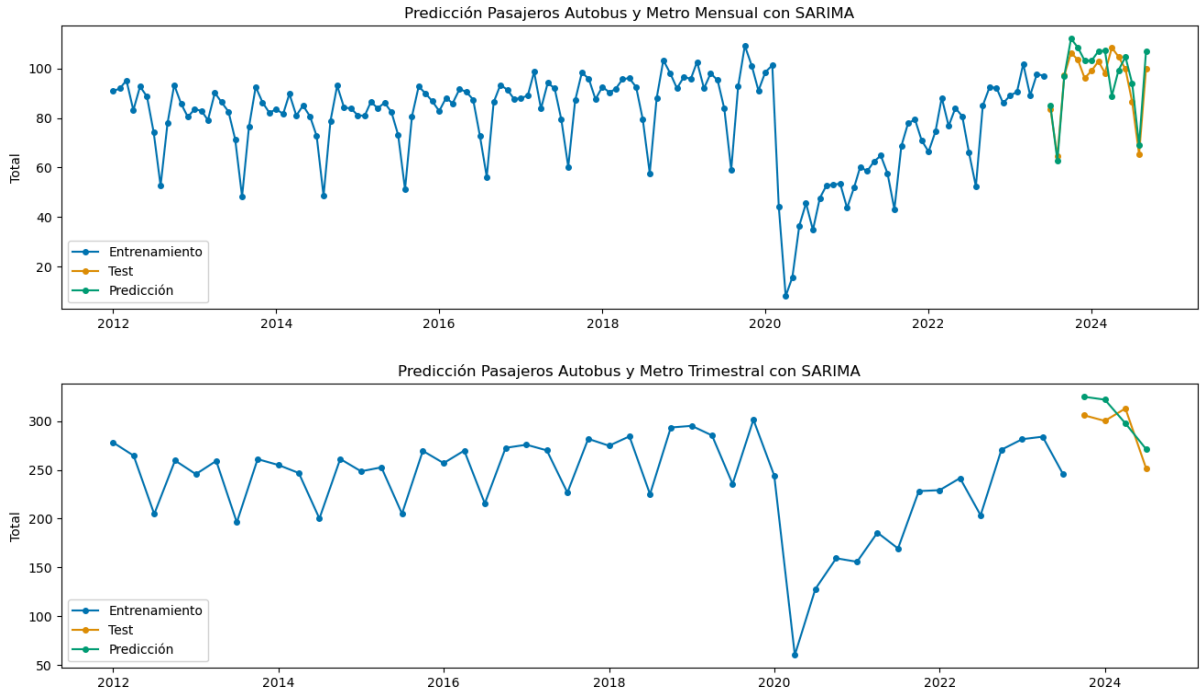
He de mencionar que no mostraré todos los modelos evaluados, pues los resultados de algunos de ellos no son nada satisfactorios. Me centraré en los modelos que mejores predicciones proporcionaron y en aquellos que aunque no dieron resultados muy buenos, tampoco son extremadamente deficientes. El lector puede comprobar los resultados de aquellos modelos que aquí no se mencionan leyendo los notebooks, los cuales están correctamente redactados y comentados.

A su vez, aclarar que he seleccionado 15 (meses) observaciones para test y predicción en el caso de las series temporales mensuales y 4 (trimestres) observaciones para test y predicción en el caso de las series trimestrales.

2. Modelos ARIMA y SARIMA

Los modelos **ARIMA** (Autoregressive Integrated Moving Average) son técnicas estadísticas para el modelado y predicción de series temporales. ARIMA combina tres componentes: la autoregresión (AR), que modela la dependencia entre observaciones pasadas; la diferenciación (I), que transforma la serie en estacionaria eliminando tendencias; y el promedio móvil (MA), que captura relaciones entre los errores residuales en diferentes momentos. Por su parte, **SARIMA** (Seasonal ARIMA) extiende ARIMA para incluir estacionalidad, añadiendo parámetros adicionales que modelan patrones cíclicos en la serie. SARIMA es especialmente útil para series temporales con componentes estacionales claras, permitiendo un ajuste más preciso en escenarios con fluctuaciones periódicas.

En nuestro caso, debido al gran componente estacional que presenta la serie temporal, el modelo SARIMA obtuvo significativamente mejores predicciones que el modelo ARIMA, tanto en las series mensuales como en las series trimestrales.



En este caso se utilizó un modelo **SARIMA (1, 1, 1)** el cual es una extensión de ARIMA que incluye estacionalidad. En este caso, utiliza un término autorregresivo (AR) de orden 1, una diferenciación de orden 1 para hacer la serie estacionaria, y un término de promedio móvil (MA) de orden 1. Se aplica además un componente estacional adicional, definido por parámetros específicos para capturar patrones cíclicos.

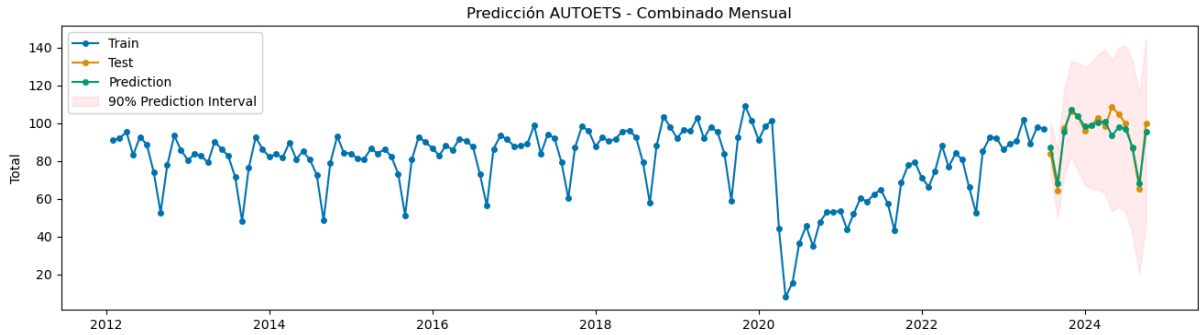
Vemos como en la serie mensual, la predicción en los primeros y en los últimos valores del test se ajusta muy bien, sin embargo, en los valores centrales difieren un poco. En la serie trimestral parece captar bien el componente estacional pero no se ajusta muy bien a la tendencia. En ambos casos el **MAPE** (Mean Absolute Percentage Error) es del 6%.

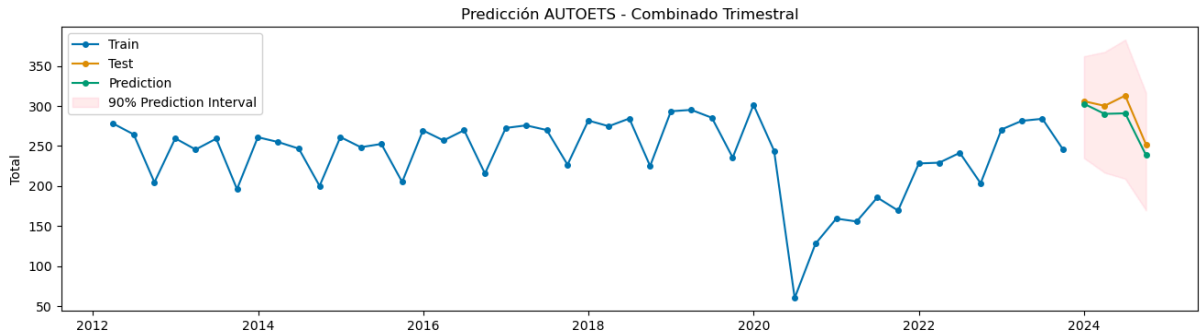
3. Modelos ETS

Los modelos **ETS** (Error-Trend-Seasonality) son una familia de modelos estadísticos diseñados para predecir series temporales mediante la descomposición en tres componentes principales: error (E), tendencia (T) y estacionalidad (S). Cada componente puede combinarse de manera aditiva o multiplicativa, dependiendo de las características de la serie. El componente de error representa las variaciones no explicadas, la tendencia captura los cambios a largo plazo, y la estacionalidad refleja patrones periódicos regulares. Estos modelos utilizan ecuaciones recursivas y suavización exponencial para actualizar las

estimaciones de nivel, tendencia y estacionalidad a medida que se incorporan nuevos datos, haciéndolos especialmente efectivos en series con patrones claros y predecibles.

En nuestro caso, estos modelos junto con los modelos Theta, ofrecieron las mejores predicciones de las series temporales. Tanto visualmente como en las métricas de error. Mientras analice las predicciones de cada modelo sólo comentaré el MAPE. En el punto de las predicciones finales, mostraré todas las métricas de error evaluadas de ambos modelos y argumentaré la elección de ETS frente a Theta para las predicciones finales.





Para la predicción de las series temporales con los modelos ETS se utilizó un enfoque AutoETS, el cual selecciona automáticamente el mejor modelo ETS para una serie temporal al evaluar combinaciones de los componentes de error (E), tendencia (T) y estacionalidad (S), ya sea aditivas o multiplicativas. Este proceso optimiza los parámetros del modelo mediante algoritmos que minimizan el error de predicción. En este caso, se pueden observar unas predicciones muy buenas en ambos casos, siendo notoriamente mejores las predicciones mensuales.

El error **MAPE** para la serie mensual fue de 3,51% y para la serie trimestral de 4,18%.

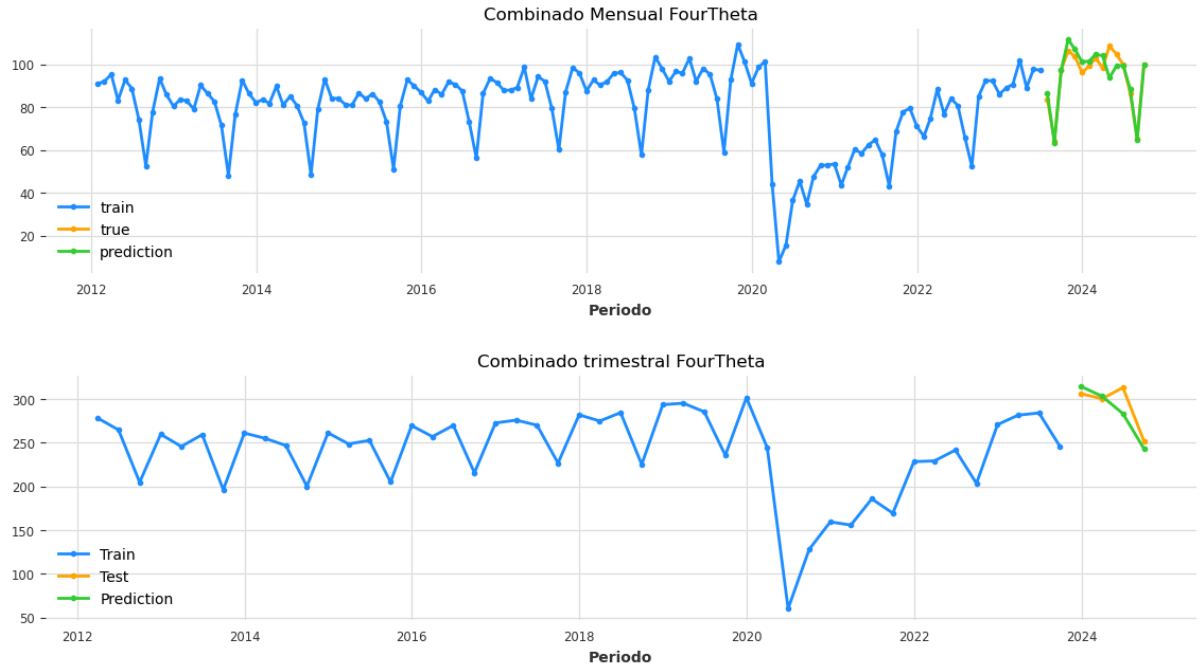
4. Modelos Theta y FourTheta

Los modelos **Theta** son una técnica estadística para la predicción de series temporales que transforman la segunda derivada de la serie mediante un parámetro denominado θ (theta), permitiendo modificar la curvatura y resaltar componentes como tendencia y estacionalidad.

El modelo descompone la serie original en diferentes líneas Theta, que se ajustan con métodos simples, como la suavización exponencial o la regresión lineal, para capturar patrones de largo plazo (tendencia) y corto plazo (variaciones estacionales). Luego, combina estas líneas para generar predicciones robustas y precisas. Es un modelo sencillo pero efectivo, especialmente en series suaves y con estacionalidad regular.

Por otro lado, el modelo **FourTheta** es una extensión del modelo Theta que incorpora elementos avanzados para mejorar su capacidad predictiva. Este modelo combina múltiples transformaciones Theta y ajustes adaptativos que se adaptan mejor a series con patrones más complejos o no lineales. FourTheta mantiene la simplicidad conceptual del modelo original, pero ofrece mayor flexibilidad y precisión, siendo útil en escenarios donde la serie presenta cambios más dinámicos o irregularidades que un modelo Theta tradicional podría no capturar.

En este caso, se aplicaron ambos modelos con la librería darts, ambos dieron resultados similares por lo que veremos los resultados del modelo Fourtheta.

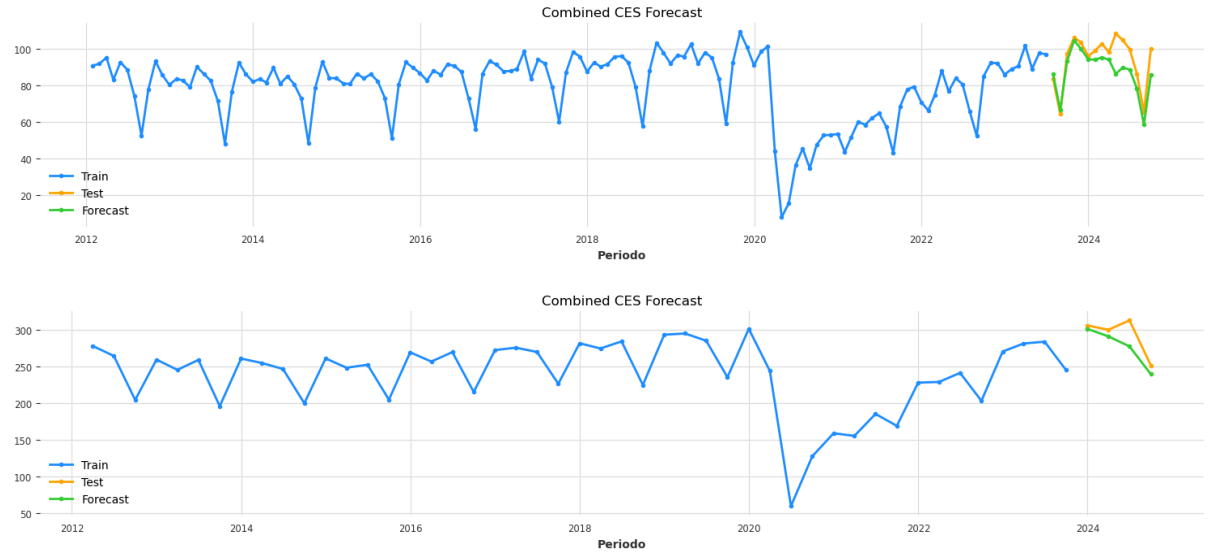


Gracias a la aplicación de técnicas de ajuste de hipérparámetros y validación cruzada la cuál proporciona una estimación más confiable de la capacidad predictiva del modelo al medir su desempeño en diferentes segmentos de la serie, hemos obtenido mejores resultados predictivos. Esto ayuda a identificar patrones generales y evita el sobreajuste, asegurando que el modelo generalice bien a datos no vistos. Además, permite comparar de manera objetiva distintos modelos o configuraciones, facilitando la selección del enfoque más adecuado para capturar las características específicas de la serie temporal.

El GridsearchCV seleccionó un **modelo aditivo**, una **estacionalidad multiplicativa** y una **tendencia lineal**.
En este caso el **MAPE** de la serie mensual fue de 3,38% y el de la serie trimestral 4,15%.

5. Modelo CES

El modelo **CES** (Compound Error, Trend, and Seasonality) es un enfoque estadístico para series temporales que combina los componentes de error, tendencia y estacionalidad de manera no descomponible, a diferencia de otros modelos como ETS. En lugar de separar los componentes, CES los modela como una única ecuación, permitiendo capturar interacciones complejas entre ellos. Es especialmente útil para series temporales con patrones regulares o estacionales, ya que puede adaptarse tanto a estructuras aditivas como multiplicativas. Este modelo destaca por su flexibilidad y capacidad para generar predicciones robustas sin requerir una descomposición explícita de los componentes de la serie.



Se observa como los puntos de las predicciones del modelo CES se ajustan bien en los valores iniciales pero no logra captar bien la tendencia creciente de los datos. Se utilizó el parámetro **model='z'** el cual indica que se seleccionará automáticamente el mejor tipo de modelo CES para la serie temporal. Este enfoque permite que el algoritmo evalúe y elija entre las variantes aditivas o multiplicativas de los componentes de tendencia y estacionalidad, optimizando la configuración para ajustar mejor los datos disponibles.

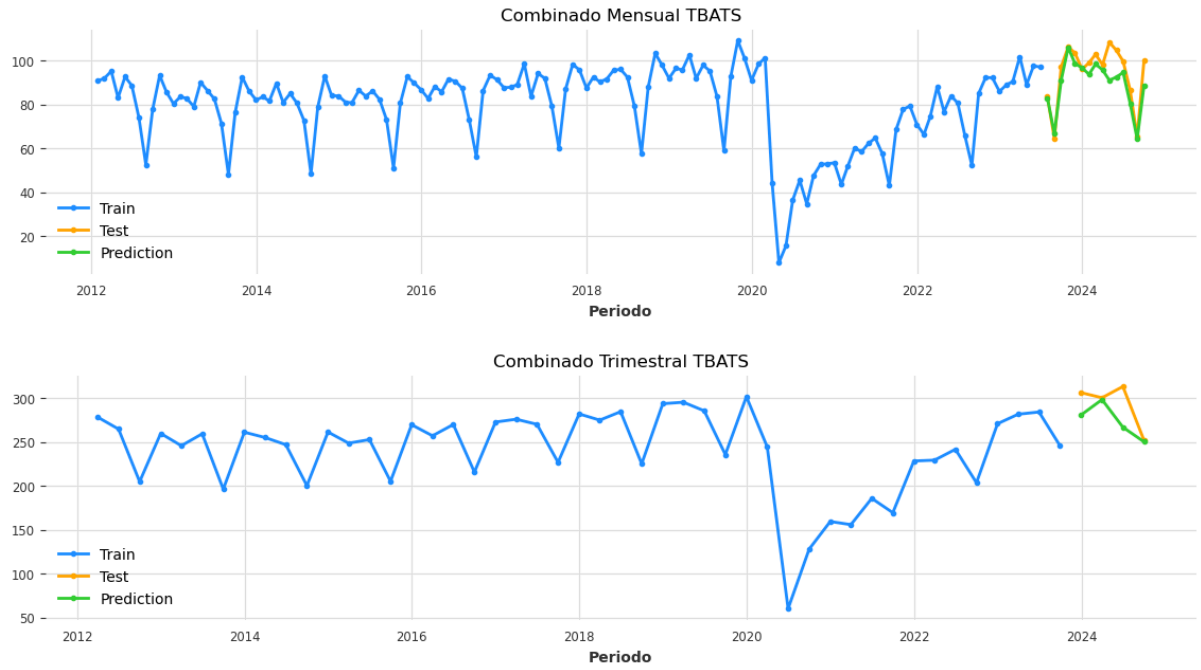
En este caso, el **MAPE** para la serie mensual fue 7,54% y para la serie trimestral 5,15%.

6. Modelos BATS y TBATS

Los modelos **BATS** (Box-Cox transformation, ARMA errors, Trend, Seasonal components) son una técnica avanzada para el modelado de series temporales, diseñada para manejar datos con patrones complejos, incluyendo múltiples estacionalidades y efectos de tendencia. Utilizan una transformación de Box-Cox para estabilizar la varianza, incorporan errores ARMA para capturar dependencias residuales, y permiten modelar tendencias lineales o no lineales junto con componentes estacionales simples o múltiples. Estos modelos son

especialmente efectivos para series temporales con estacionalidad compleja, como las que tienen patrones anuales y semanales simultáneamente.

El modelo **TBATS** (Trigonometric BATS) es una extensión del modelo BATS, diseñada específicamente para manejar estacionalidades largas y complejas utilizando funciones trigonométricas para representar los componentes estacionales. Esta característica lo hace particularmente útil para datos con estacionalidades no regulares o muy largas, como patrones mensuales en series de varios años. Al igual que BATS, TBATS utiliza transformaciones de Box-Cox, errores ARMA y componentes de tendencia, pero destaca por su capacidad de modelar de forma eficiente y precisa estacionalidades que otros enfoques podrían no capturar adecuadamente.



En nuestro caso, los modelos TBATS se comportaron mejor que los modelos BATS y generaron predicciones ligeramente superiores. Se utilizó transformación Box-Cox para estabilizar las varianzas y mejorar la normalidad de las series temporales, además de indicar que la serie temporal presenta un componente de tendencia ligeramente positivo.

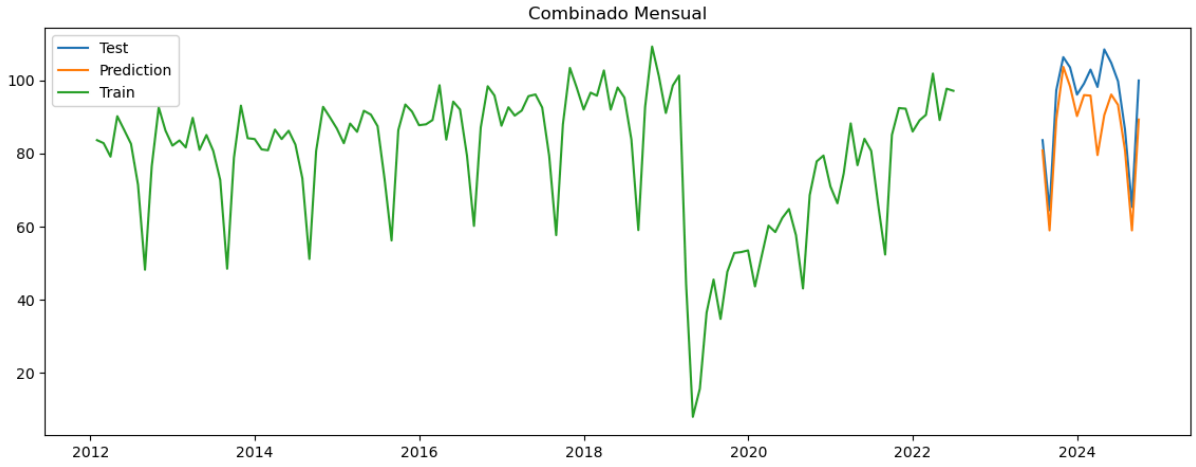
En este caso, el **MAPE** para la serie mensual fue de 5,39% y el de la serie trimestral de 6,11%.

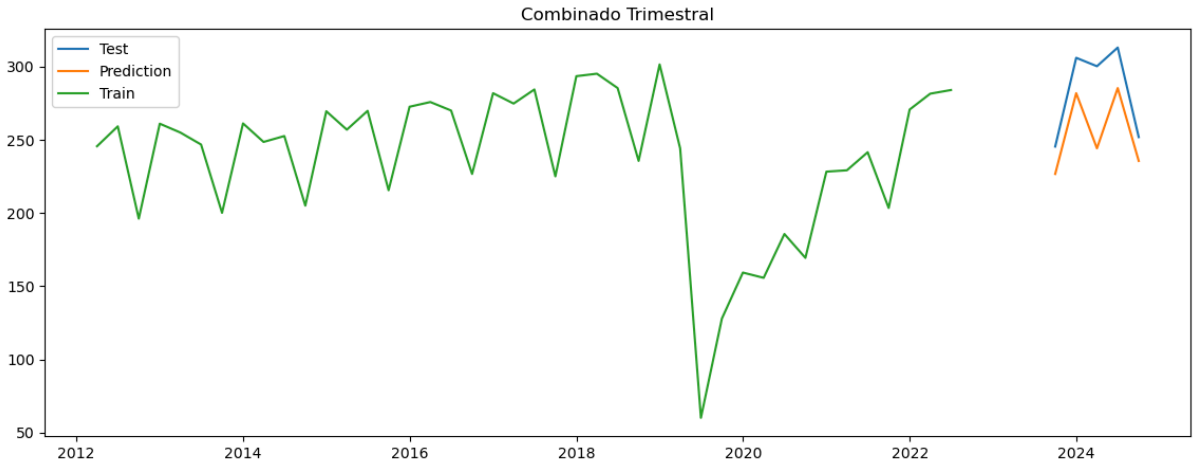
7. Modelo KNN (K-Nearest Neighbors)

El modelo **KNN** (K-Nearest Neighbors) es un algoritmo de aprendizaje supervisado utilizado tanto para clasificación como para regresión, que basa sus predicciones en la similitud entre

observaciones. En el contexto de series temporales, KNN predice el valor futuro identificando los k puntos más cercanos en el historial de datos según una medida de distancia (como Euclidiana) y promediando sus valores. Este modelo no asume una estructura específica en los datos, lo que lo hace flexible, pero depende en gran medida de la selección del número de vecinos (k) y de la calidad de los datos para obtener resultados precisos.

En nuestro caso, de todos los modelos de ML (Machine Learning) aplicados, este modelo fue el que proporcionó los mejores resultados. Los modelos restantes no se ajustaron nada bien a los datos a la hora de generar predicciones.





Como se puede observar visualmente, aunque las predicciones mensuales no son del todo erróneas, hemos observado a lo largo de todo este informe otros modelos que se ajustan mejor a este tipo de series temporales.

Hemos comprobado que los modelos de ML no son completamente adecuados para series temporales porque no consideran explícitamente las dependencias temporales entre observaciones, lo que puede llevar a predicciones menos precisas si no se capturan correctamente las tendencias y estacionalidades inherentes. Además, requieren características adicionales derivadas como rezagos (lags) para modelar patrones temporales, lo que aumenta la complejidad. Finalmente, suelen ser menos interpretables que los modelos estadísticos

tradicionales, dificultando el análisis de los resultados en contextos donde la comprensión de los patrones es clave.

En el caso del modelo KNN, el **MAPE** para la serie temporal mensual fue 8,07% y para la serie trimestral 9,90%.

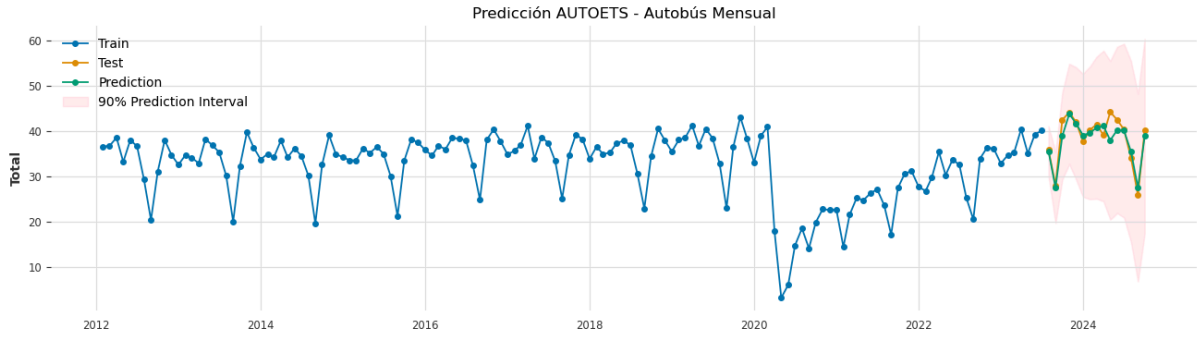
8. Modelo Final

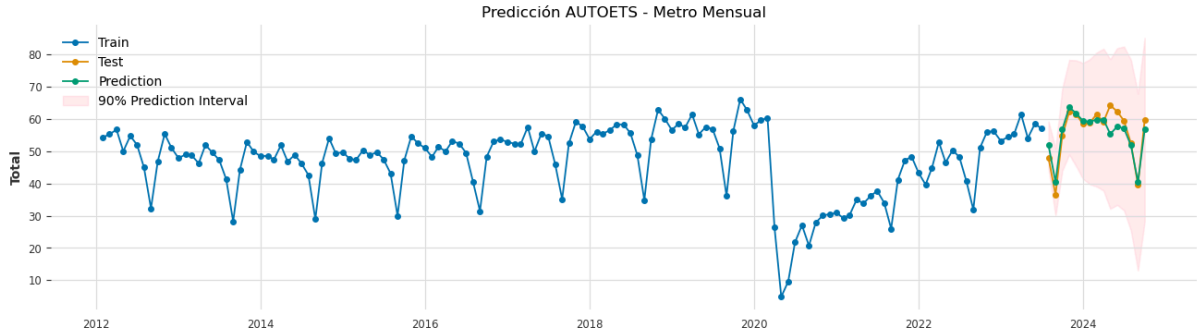
A lo largo de este estudio, se llevó a cabo un exhaustivo análisis de diversos modelos de predicción aplicados a la serie temporal de pasajeros del transporte público de Madrid. Se probaron y evaluaron tanto modelos estadísticos clásicos (ARIMA, ETS, Theta, CES, BATS, TBATS, Prophet, BSTS) como modelos avanzados de machine learning (Decision Tree, XGBoost, LightGBM, Random Forest y KNN). Cada modelo fue ajustado utilizando técnicas de ajuste de hiperparámetros y validación cruzada, con el objetivo de maximizar su capacidad predictiva y asegurar un enfoque riguroso y reproducible.

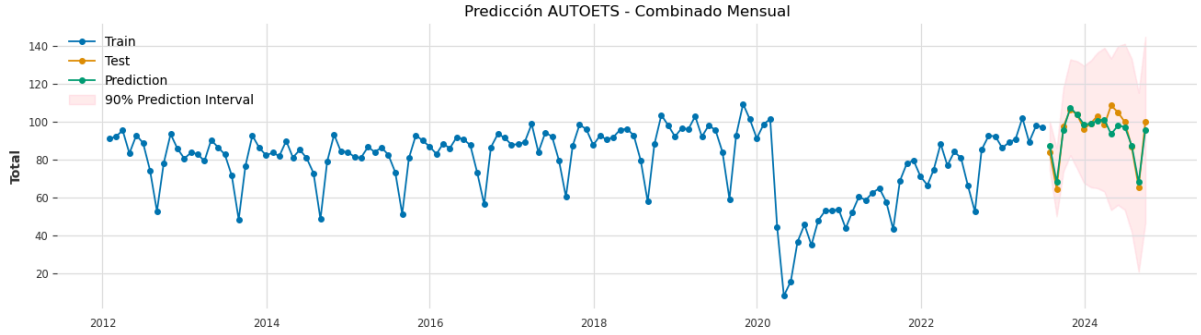
La evaluación se realizó mediante el graficado de las predicciones con sus intervalos de confianza y el cálculo de métricas clave: el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE) y el Error Absoluto Medio Porcentual (MAPE). Estas métricas permitieron comparar el desempeño de los modelos de manera objetiva y cuantitativa. Si bien varios modelos presentaron resultados competitivos, los modelos ETS y Theta destacaron significativamente por su capacidad de capturar los patrones de tendencia y estacionalidad inherentes a los datos de transporte público.

8.1 Predicciones con ETS

Series mensuales



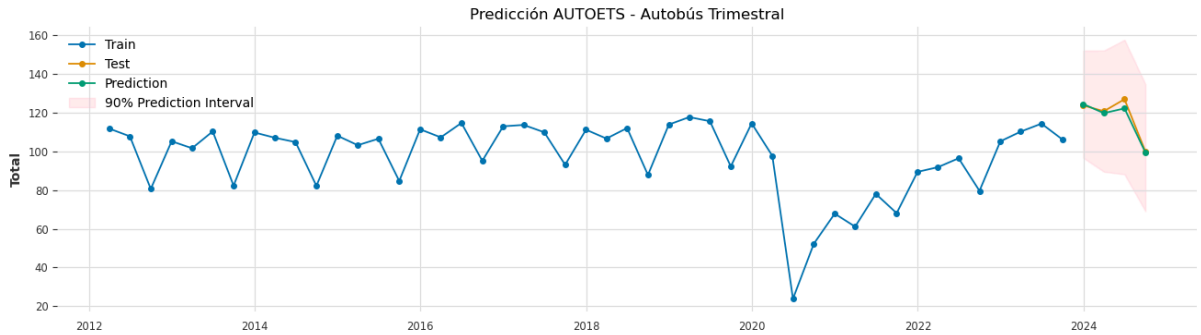


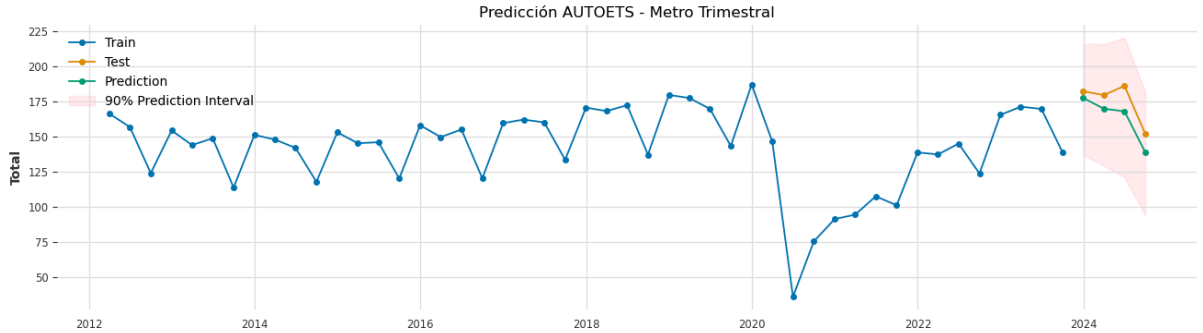


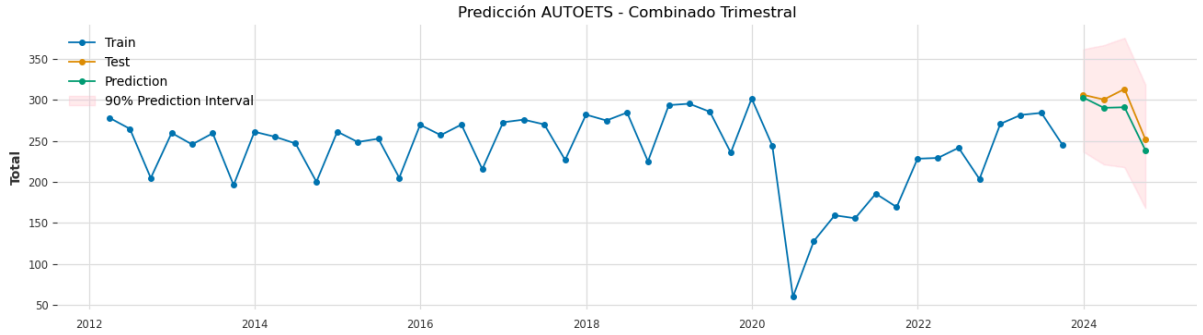

```
Métricas para Autobús Mensual:
MSE: 4.718598852179303, RMSE: 2.1722336090253513, MAPE: 0.038964134894091666

Métricas para Metro Mensual:
MSE: 10.574237900366414, RMSE: 3.251805329408022, MAPE: 0.04355536794726542

Métricas para Combinado Mensual:
MSE: 23.827076473422597, RMSE: 4.881298646202934, MAPE: 0.03511846408082429
```

Observamos unas métricas muy buenas en conjunto para las predicciones mensuales con el modelo ETS.

Series trimestrales





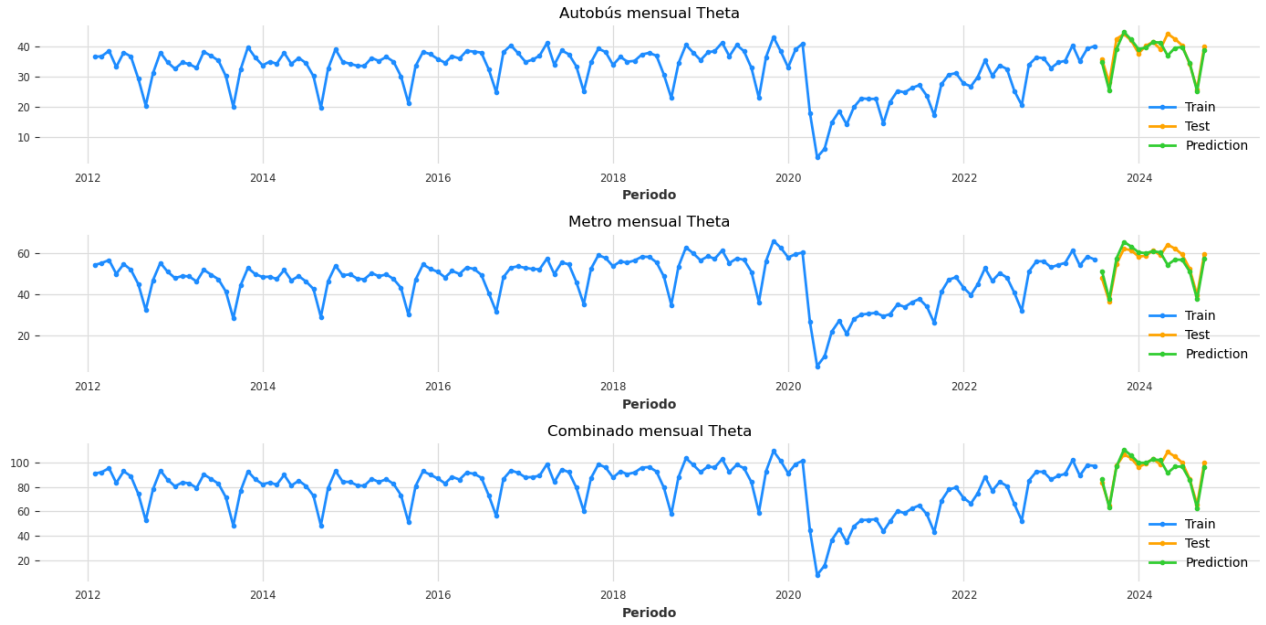



```
Métricas para Autobús Trimestral:
MSE: 6.041031249196908, RMSE: 2.4578509412079708, MAPE: 0.014211584720710659

Métricas para Metro Trimestral:
MSE: 153.83868163959468, RMSE: 12.40317224098717, MAPE: 0.06590390395338475

Métricas para Combinado Trimestral:
MSE: 194.2642574724126, RMSE: 13.937871339354967, MAPE: 0.041847361695736864
```

Observamos buenas métricas en conjunto también para las predicciones trimestrales, en especial para la serie trimestral de autobús el modelo ETS es el que mejor predicción ha generado con un MAPE de 1,41%.

8.2 Predicciones con Theta

Series mensuales




```
Métricas para Autobús Mensual:
MSE: 6.111005318196414, RMSE: 2.472044764602052, MAPE: 4.3804097426450745

Métricas para Metro Mensual:
MSE: 12.520550345210069, RMSE: 3.5384389701123955, MAPE: 4.8271719731139155

Métricas para Combinado Mensual:
MSE: 29.8722117232718, RMSE: 5.465547705699018, MAPE: 3.707616743801179
```

Se observan buenas métricas de error para las predicciones de las series mensuales con el modelo Theta, aunque las métricas del modelo ETS son ligeramente superiores.

Series trimestrales


```
Métricas para Autobús Trimestral:
MSE: 32.609415789087436, RMSE: 5.710465461684138, MAPE: 4.53129553312913

Métricas para Metro Trimestral:
MSE: 186.5564606272127, RMSE: 13.658567297751718, MAPE: 5.845973038968093

Métricas para Combinado Trimestral:
MSE: 259.47263135027555, RMSE: 16.10815418818294, MAPE: 4.156256524555424
```

Tras un análisis exhaustivo de los modelos ETS y Theta aplicados a la predicción de pasajeros del transporte público de Madrid, se llevó a cabo una evaluación detallada tanto visual como cuantitativa. En el aspecto visual, las predicciones del modelo ETS mostraron un ajuste más preciso a los datos históricos, capturando con mayor fidelidad los patrones de tendencia y estacionalidad característicos de la demanda de transporte en la ciudad, sobre todo en las series trimestrales.

Desde el punto de vista de las métricas, el modelo ETS presentó ligeramente mejores resultados en el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE) y el Error Absoluto Medio Porcentual (MAPE). Estas métricas indican una menor desviación entre los valores reales y los predichos, lo que sugiere una mayor precisión y robustez en las predicciones del modelo ETS frente al modelo Theta.

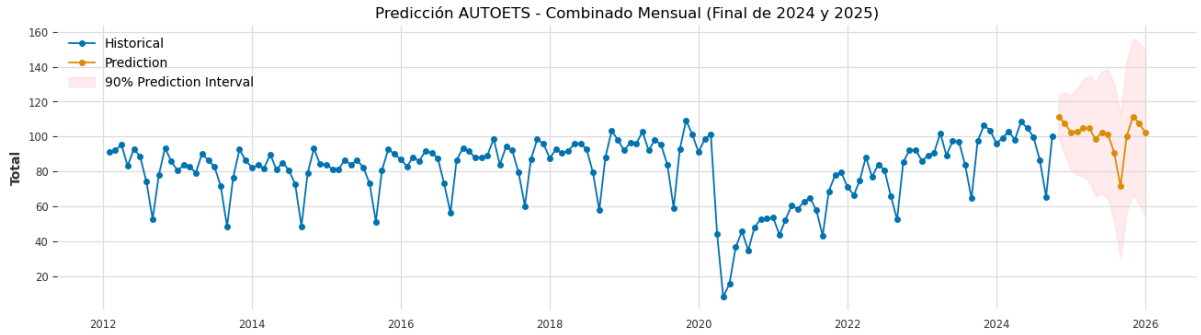
Considerando tanto el desempeño cuantitativo como cualitativo, se ha determinado que el modelo ETS es el más adecuado para este problema. Por ello, se selecciona como el modelo definitivo para realizar las predicciones finales, asegurando una mayor confiabilidad en la estimación de la demanda futura de pasajeros del sistema de transporte público de Madrid. Este enfoque permitirá respaldar decisiones informadas en la planificación y gestión del transporte urbano.

9. Predicciones Finales

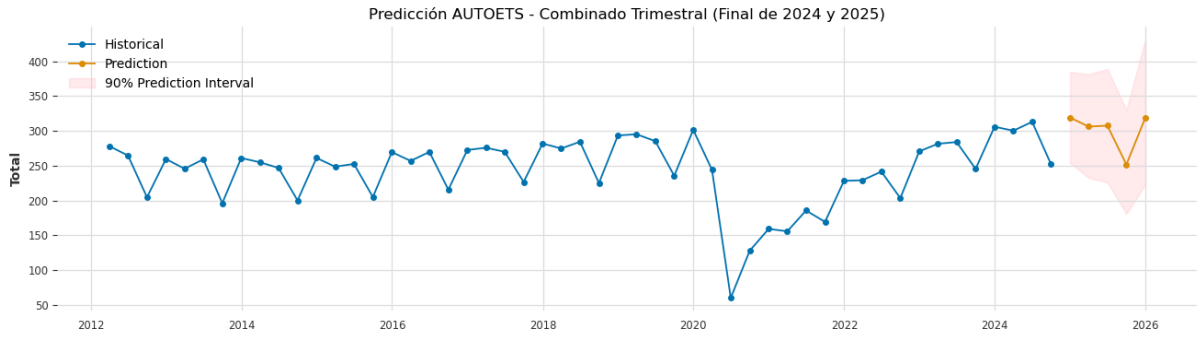
Tras este análisis exhaustivo, a continuación se presentan las predicciones finales para la serie temporal de transporte mensual y trimestral.

Esta predicción comprende un horizonte temporal de 15 (meses) observaciones para las series temporales mensuales y 5 (trimestres) observaciones para las series mensuales trimestrales. Ambos periodos comprenden el final del año 2024 y todo el año 2025.

Predicción mensual



Predicción trimestral



Ambas predicciones se visualizan de manera precisa y consistente, observando como el modelo capta los patrones de tendencia y estacionalidad.

A continuación se redactará el plan de acción económico en base a las predicciones generadas.

10. Plan de Acción Económico

A partir de las predicciones generadas mediante el modelo AutoETS para los pasajeros del transporte público de Madrid (mensual y trimestral), se plantea un plan de acción centrado en la eficiencia operativa, la satisfacción del usuario y la sostenibilidad. Este enfoque busca optimizar recursos, mejorar la experiencia de los usuarios y garantizar la sostenibilidad económica y ambiental del sistema de transporte. Los beneficios de este proyecto destacan la

importancia de integrar predicciones basadas en datos para una toma de decisiones estratégica y efectiva.

10.1 Eficiencia Operativa

Para reducir costes y maximizar la eficiencia del sistema de transporte público, se propone:

- **Optimización de horarios y frecuencias**: Ajustar la frecuencia de autobuses y trenes en función de los niveles de demanda previstos por las predicciones. Durante periodos de menor actividad como hemos visto en el tercer trimestre del año, donde la actividad económica de Madrid es menor y los centros educativos están cerrados, reducir la frecuencia para evitar costes operativos innecesarios.
- **Reasignación de recursos**: Dirigir vehículos adicionales hacia líneas con picos de demanda previstos, asegurando una distribución equitativa y eficiente de los recursos.
- **Uso de tecnología**: Implementar sistemas de monitoreo en tiempo real mediante IoT para detectar y responder rápidamente a fluctuaciones inesperadas en la demanda.
- **Revisión de contratos**: Negociar contratos con proveedores de energía y mantenimiento basados en los patrones de uso previstos, reduciendo costes fijos.

10.2 Satisfacción del Usuario

La predicción precisa de la demanda permite implementar mejoras en la calidad del servicio para aumentar la satisfacción de los usuarios:

- **Refuerzos en picos de demanda**: Incrementar la frecuencia de vehículos en horarios y días clave según las predicciones, evitando aglomeraciones y tiempos de espera prolongados.
- **Comunicación efectiva**: Desarrollar aplicaciones móviles o paneles informativos que brinden a los usuarios datos en tiempo real sobre la disponibilidad de vehículos, tiempos de espera y niveles de ocupación.
- **Incentivos durante horas valle**: Ofrecer descuentos o pases especiales en horarios de menor demanda para incentivar el uso del transporte en periodos menos concurridos.
- **Encuestas continuas**: Implementar encuestas regulares para monitorear la percepción del usuario y adaptar las estrategias en función de sus necesidades.

10.3 Sostenibilidad

La alineación de recursos con las necesidades reales contribuye significativamente a la sostenibilidad económica y ambiental:

- **Transición hacia flotas ecológicas**: Priorizar la inversión en autobuses eléctricos o de bajas emisiones para las líneas con mayor demanda, maximizando su impacto positivo en el medio ambiente.
- **Reducción de emisiones**: Ajustar los servicios en horas valle para disminuir el consumo energético y reducir las emisiones de gases contaminantes.

- **Promoción del transporte público**: Diseñar campañas que resalten los beneficios ambientales y económicos del uso del transporte público frente a vehículos privados.
- **Eficiencia en el uso de energía**: Implementar tecnologías que optimicen el consumo de energía en estaciones, trenes y autobuses, como sistemas de iluminación eficiente y recuperación de energía en frenado.

10.4 Importancia del Proyecto de Predicción

Un proyecto de predicción de series temporales, como el desarrollado aquí, no solo permite anticipar escenarios futuros, sino también fundamentar decisiones estratégicas con datos objetivos. En el contexto del transporte público, estas herramientas son esenciales para:

1. **Planificación eficiente**: Permiten adaptar los servicios a las necesidades reales, reduciendo costes y mejorando la experiencia del usuario.
2. **Competitividad y sostenibilidad**: Garantizan un sistema más competitivo y sostenible frente a los retos económicos y medioambientales.
3. **Impacto social y económico**: Al optimizar el transporte público, se fomenta una mayor inclusión social, reduciendo la dependencia del vehículo privado y mejorando la calidad de vida de los ciudadanos.

Este plan de acción demuestra la importancia de utilizar datos predictivos para transformar la gestión del transporte público, alineando los recursos con la demanda, mejorando la calidad del servicio y garantizando un futuro más sostenible para la Comunidad de Madrid.

Conclusión

En este informe, hemos analizado y evaluado diferentes enfoques para la predicción de la demanda de pasajeros de metro y autobuses en Madrid, partiendo de un dataset compuesto por la fecha y la demanda total de pasajeros. Los resultados indican que los enfoques más apropiados para este tipo de series temporales son los modelos univariados, como ETS, Theta, que permiten capturar de manera precisa patrones estacionales, tendencias y dinámicas intrínsecas de la demanda.

Este análisis proporciona una base sólida para futuras investigaciones y mejoras en las predicciones, destacando la importancia de disponer de datos adicionales o variables exógenas que enriquezcan el modelo y permitan incorporar factores contextuales que influyen en la demanda. Así, se podrá optimizar la planificación y gestión del transporte público en Madrid, contribuyendo a un servicio más eficiente y adaptado a las necesidades de los usuarios.

Anexo I

Con el objetivo de proporcionar una mayor comprensión al informe, a continuación se redactarán cuestiones cuyo objetivo es resolver posibles dudas o preguntas que pudieran surgir al lector durante la lectura del mismo.

Metodología técnica

En cuanto a la metodología utilizada para la evaluación técnica de los modelos se han priorizado las métricas ya comentadas anteriormente las cuáles serán definidas a continuación:

- **MSE (Mean Squared Error):**
  Es la media de los errores al cuadrado entre los valores reales y las predicciones. Penaliza con más fuerza los errores grandes, lo que lo hace útil cuando se quiere evitar grandes desviaciones en las predicciones.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- **RMSE (Root Mean Squared Error):**
  Es la raíz cuadrada del MSE, lo que devuelve la métrica a las mismas unidades que los datos originales. Es fácil de interpretar y sensible a los errores grandes, al igual que el MSE.

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

- **MAPE (Mean Absolute Percentage Error):**
  Mide el error porcentual promedio entre los valores reales y las predicciones, lo que permite interpretar el error relativo al tamaño de los datos. Es útil para comparaciones entre diferentes escalas de datos.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - p_i}{y_i}\right|$$

Estas métricas permiten evaluar la precisión del modelo de predicción desde diferentes perspectivas, adaptándose a las necesidades específicas del análisis de la demanda de transporte público:

- **MSE y RMSE** destacan los errores grandes, lo cual es crucial al predecir la demanda de metro y autobuses, ya que desviaciones significativas pueden llevar a problemas como la falta de capacidad o recursos excedentes, impactando la experiencia del usuario y la eficiencia operativa.
- **MAPE** es particularmente útil para interpretar el rendimiento relativo del modelo, ya que normaliza los errores en función de la demanda real. Esto es relevante porque la magnitud de la demanda puede variar significativamente entre meses o estaciones del año.

Al combinar estas métricas, se obtiene un análisis más robusto que facilita una mejor planificación y toma de decisiones, asegurando que las predicciones sean precisas tanto en términos absolutos como relativos.

Modelos No Utilizados

En cuanto a la utilización de modelos para realizar las predicciones se han descartado el uso de los siguientes modelos derivados de las características inherentes a los mismos las cuáles no se ajustaban de una manera correcta a los datos disponibles.

Modelos de series jerárquicas

Los modelos de series jerárquicas son aquellos que estructuran y analizan series temporales en niveles organizados jerárquicamente, donde las predicciones en niveles superiores (como totales agregados) deben ser consistentes con las de niveles inferiores (como subgrupos o componentes específicos). Este enfoque es útil cuando los datos están desglosados en varias categorías o regiones y existe interés en garantizar coherencia entre las predicciones de diferentes niveles.

Dado que el dataset contiene únicamente dos columnas (fecha y demanda total de pasajeros), los modelos de series jerárquicas no son apropiados, ya que están diseñados para trabajar con datos desagregados en niveles jerárquicos. En este caso, no hay subniveles o categorías que justifiquen su uso. Además, estos modelos no aprovechan las características temporales o exógenas que podrían influir en la demanda, como variaciones estacionales, eventos especiales o factores climáticos, lo que limita su capacidad para capturar patrones relevantes en los datos disponibles.

Modelo Causal Impact

El modelo causal Causal Impact es una herramienta que utiliza métodos bayesianos para medir el efecto de un evento o intervención en una serie temporal. Se basa en construir un modelo predictivo de la serie en ausencia de la intervención y luego comparar las predicciones con los valores reales tras el evento. Este enfoque es ideal para identificar cambios causados por intervenciones específicas, como políticas, cambios en infraestructura o eventos disruptivos.

Dado que el dataset solo incluye la fecha y la demanda total de pasajeros, el modelo Causal Impact no es apropiado porque está diseñado para evaluar el impacto de eventos o

intervenciones específicas, lo que requiere información adicional sobre cuándo y qué eventos ocurrieron. Sin estas variables exógenas o contextuales que expliquen posibles cambios en la demanda, el modelo no puede distinguir entre fluctuaciones naturales de la serie y efectos causales. Además, el enfoque se centra en análisis retrospectivos de impacto, más que en realizar predicciones precisas de la demanda futura, que es el objetivo principal en este caso. Por ello, sería más adecuado utilizar modelos puramente temporales que capturen patrones intrínsecos de la serie.

Modelo ARCH

Los modelos ARCH (Autoregressive Conditional Heteroskedasticity) son herramientas estadísticas diseñadas para modelar y predecir la variabilidad (volatilidad) de una serie temporal. Son especialmente útiles cuando la varianza de los datos no es constante a lo largo del tiempo (heterocedasticidad condicional). Estos modelos se emplean comúnmente en series financieras, donde las fluctuaciones en la volatilidad son significativas y afectan la dinámica de los datos.

El modelo ARCH no es adecuado para predecir la demanda de pasajeros en metro y autobuses en Madrid, ya que este tipo de series temporales generalmente se caracteriza por patrones de estacionalidad y tendencia, más que por cambios significativos en la varianza condicional a lo largo del tiempo. Dado que el dataset solo incluye la fecha y la demanda total de pasajeros, los modelos ARCH no pueden capturar patrones relevantes como fluctuaciones diarias, semanales o estacionales. Además, su enfoque en modelar la variabilidad es menos útil para predecir valores absolutos de demanda, que es el objetivo principal en este caso.

Modelo VAR

El modelo VAR (Vector Autoregressive) es un enfoque estadístico que captura las interrelaciones entre múltiples series temporales. Cada variable se modela como una función lineal de sus valores pasados y de los valores pasados de otras variables incluidas en el modelo. Este enfoque es ideal para analizar y predecir sistemas dinámicos donde varias series están interconectadas.

El modelo VAR no es adecuado para predecir la demanda de pasajeros de metro y autobuses en Madrid porque requiere al menos dos o más series temporales relacionadas para capturar las interacciones entre ellas. Dado que el dataset solo contiene una columna de fecha y otra de demanda total de pasajeros, no hay variables adicionales que justifiquen el uso de un modelo multivariado como VAR. Además, VAR no está diseñado para manejar patrones estacionales o de tendencia comunes en datos de transporte, a menos que se apliquen transformaciones adicionales.

Modelo HMM

Los HMM (Hidden Markov Models) son modelos probabilísticos que asumen que una serie temporal está gobernada por un conjunto de estados ocultos no observables. Estos estados evolucionan a lo largo del tiempo siguiendo un proceso de Markov, y las observaciones son

generadas en función del estado actual. Son útiles para analizar series con cambios estructurales o patrones que dependen de un régimen subyacente, como ciclos económicos o sistemas con múltiples fases.

El modelo HMM no es adecuado para predecir la demanda de pasajeros en metro y autobuses en Madrid porque está diseñado para identificar patrones ocultos en series temporales con múltiples estados subyacentes, lo cual no es relevante si el dataset solo incluye la fecha y la demanda total de pasajeros. Sin información adicional sobre posibles cambios de régimen (como alteraciones en el sistema de transporte o eventos disruptivos), el modelo no puede aportar ventajas significativas. Además, HMM requiere una gran cantidad de supuestos y parametrización, lo que puede resultar en un ajuste excesivo y predicciones menos precisas.